UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works

Flavio M. Cecchini, Rachele Sprugnoli, Giovanni Moretti, Marco Passarotti
CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milano
{flavio.cecchini,rachele.sprugnoli,
giovanni.moretti,marco.passarotti}@unicatt.it

Abstract

English. This paper[1] presents the early stages of the development of a new treebank containing all of Dante Alighieri's Latin works. In particular, it describes the conversion of the original TEI-XML files to CoNLL-U, the creation of a gold standard, the process of training four annotators and the evaluation of the syntactic annotation in terms of inter-annotator agreement and LA, UAS and LAS. The aim is to release a new resource, in view of the celebrations for the 700th anniversary of Dante's death, which can support the development of the *Vocabolario Dantesco*.

1 Introduction

The research field of treebanking (i. e. the building of corpora enhanced with syntactic metadata) has evolved substantially since the time when the first large-scale syntactically annotated corpus, the Penn Treebank for English, was published between the late Eighties and the early Nineties (Taylor et al., 2003). Across the last two decades, the range of languages for which a treebank is available has increased considerably. The grammar framework behind the most widespread annotation style currently used in treebanking has also changed: treebanks annotated according to various styles of dependency grammars have been increasingly outnumbering those based on constituency (or phrase-structure) grammars, as demonstrated by the current status of the *Universal Dependencies* initiative (UD) containing more than 160 treebanks and 90 languages which follow the same, dependency-based, annotation style (Nivre et al., 2016).

The set of textual genres covered by currently available treebanks is quite diverse. While the first corpora were built mostly collecting texts from news, the last decade has seen a substantial growth of treebanks of different genres, including literary texts, mostly written in ancient or historical languages.[2] The first available treebanks for ancient languages were those for Ancient Greek and/or Latin, namely the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2019) and the Ancient Greek and Latin Dependency Treebank from the Perseus digital library (Bamman and Crane, 2011). With regard to Latin, the available treebanks in UD cover just a minimal subset of the Latin texts that have survived the centuries and which show a wide diversity, mostly due to Latin's *lingua franca* role played all over Europe up until the 1800s (Leonhardt, 2009). So far, the treebanks for Latin include only portions of the Classical and Late Latin canon of texts (Perseus and PROIEL (Eckhoff et al., 2018)), a set of Early Medieval charters from Tuscia (Late Latin Charter Treebank (Korkiakangas and Passarotti, 2011; Cecchini et al., 2020)) and a selection of Late Medieval philosophical-theological texts by Thomas Aquinas (IT-TB), for a total of more than 800 000 nodes.

Among the many Latin texts that still lack syntactic annotation are those by Dante Alighieri (1265-1321). Given the importance of Dante in the history of Italian literature (and beyond) and in the light of the celebrations for the upcoming 700th anniversary of his death, we have started a project (called *UDante*) aimed at performing a UD-compliant syntactic annotation of all his Latin texts. The syntactic annotation of Dante's *opera omnia* in Latin fits into the larger project of the *Vocabolario Dantesco*, which aspires to provide a detailed description of the entire (both Vulgar and

[2] Examples among the UD treebanks are the Kyoto treebank of Classical Chinese (Yasuoka, 2019) and the *Scriptorium* treebank of Coptic (Zeldes and Abrams, 2018).

Latin) lexicon of Dante Alighieri.[3] Indeed, during the composition of entries for the vocabulary, lexicographers will benefit from having the possibility to run syntactic queries on Dante's works.

The choice of using the UD formalism in the *UDante* project is motivated by a number of benefits implied by the inclusion of a new set of annotated texts into such a large collection of treebanks sharing the same annotation style, among others the use of the several tools developed by the UD community with the goal of querying, editing, visualizing and (automatically) processing the (meta)data of the treebanks.[4] Particularly, a remarkable added value is the possibility to run common queries on the almost 100 different languages provided with at least one treebank in the current version of UD (v2.6, released on May 15th, 2020). Furthermore, adopting a well known and widely used data format (CoNLL-U) and part-of-speech tagset (UPOS) fosters the dissemination and use of a treebank of Dante's Latin works in the community of computational linguistics, laying the foundation for a closer collaboration with that of Italian phylology and, more generally, with scholars in the Humanities, leading to a mutual benefit.

This paper presents the process behind the development of the manually annotated UD treebank containing the full collection of Latin works of Dante Alighieri. More specifically, we describe the conversion of the original TEI-XML files into the CoNLL-U format, we give details on the creation of a gold standard[5] and we report on the training of four annotators with no previous knowledge of the UD formalism, providing an evaluation of their annotation work.

2 Treebank Development

The texts of the Latin works by Dante Alighieri (*De monarchia*, *De vulgari eloquentia*, *Eclogues*, *Epistulae* and *Quaestio de aqua et terra*) are made available by the *DanteSearch* corpus (Tavoni, 2012).[6] All texts come already manually lemmatised and morphologically tagged by a team of young scholars at the University of Pisa, and are encoded in TEI-XML.[7] The original files are converted into the CoNLL-U format[8] and then revised and syntactically annotated using ConlluEditor (Heinecke, 2019).

2.1 From TEI-XML to CoNLL-U

We implement an own developed script to automatically convert the TEI-XML files of the *DanteSearch* corpus into the CoNLL-U format. First of all, the script analyses the XML tag structure to identify the internal organisation of the text (i. e. the division of the work in books, chapters etc.): this information is stored in the `MISC` field so as to facilitate the recoverability of the original structure of the text starting from the CoNLL-U file. Then, sentences are split and the tag `<LM>`, which for each token contains morphological information, is parsed in order to extract lemma, part of speech and morphological traits, and to convert the codes used in *DanteSearch* into UPOS tags[9] (originally inspired by (Petrov et al., 2012)) and UD features respectively. An example is:

```
<LM lemma="resono"
catg="va1cis3">resonaret</LM>
```

In particular, the part-of-speech tag and the morphological traits are derived from the values of the `catg` attribute, while those fields of the CoNLL-U format dedicated to syntactic information are filled with underscores (_) and left for manual annotation. The conversion of `catg` attribute is challenging, because the string-type values of its slots do not follow a fixed-position strategy, thus the string ends up having a variable length, and the same morphological trait can occupy different positions according to the given part of speech. In general, UD requires a more fine-grained annotation of morphological traits compared to the one originally provided by the *DanteSearch* corpus. For example, the value `va1cis3` of *resonaret* (active imperfect subjunctive third-person singular of *resono* 'to resound') is converted into the UD formalism as follows:

v → `VERB`
a → `Voice=Act`

[3] `http://www.vocabolariodantesco.it`

[4] See the website of UD for the list of tools: `https://universaldependencies.org`.

[5] `https://github.com/CIRCSE/UDante`.

[6] `https://dantesearch.dantenetwork.it`

[7] The *DanteSearch* corpus also contains the Vulgar works by Dante Alighieri that were used to develop a part-of-speech tagger for XIIIth century Italian by means of TreeTagger and the Stanford POS tagger (Basile and Sangati, 2016).

[8] `https://universaldependencies.org/format.html`

[9] `https://universaldependencies.org/u/pos/index.html`

1 → `VerbClass=LatA`[10]
ci → `Aspect=Imp|Mood=Sub|Tense=Past |VerbForm=Fin`
s → `Number=Sing`
3 → `Person=3`

Ad hoc rules are added to cover specific cases. For example, in *DanteSearch* the lemma *prius* 'before' is marked only with the grammatical category `r`: a rule converts `r` into the UPOS tag `ADV` and adds the morphological feature `Degree=Cmp` (comparative degree).
Annotators, in addition to annotating syntax from scratch, have to check the correctness of the automatic conversion and to manually modify or add items not covered by it. For example, annotators have to: (i) modify the grammatical category of population names (such as *Veronenses* 'inhabitants of Verona'), which are marked as proper names in *DanteSearch*, contrary to UD recommendations, for which they should be considered as adjectives;[11] (ii) check the ambiguous case of some pronouns in the neutral gender which in *DanteSearch* have mistakenly been marked as nominative instead of accusative (e. g. *quod* 'that'), or viceversa; (iii) disambiguate the `PronType` feature in the case it has more than one value: this happens because the types of pronouns in *DanteSearch* cannot always be matched to only one `PronType` value (e. g. *quis* 'who/any', interrogative or indefinite).

2.2 Gold Standard Creation

An important part of the *UDante* project consists in training a group of annotators on the formalism of UD with the goal of providing them with adequate competences to pursue the complete syntactic annotation of Dante's works. To this aim, for each of the two parts of our training (Section 2.3) a small number of sentences is singled out from all across Dante's Latin texts to be used as a common (first part) or individual (second part) benchmark for the assessment of the annotators' progress.

The first part of the training makes use of 33 sentences out of the total 1 662 (corresponding to 950 tokens out of 55 666). These sentences are not chosen to be consecutive, nor do they follow a particular order, but they are allocated into three different groups of increasing complexity, corresponding to the three distinct phases of this part of the training. The distribution is of 15 sentences in the first, introductory group, 5 in the second, intermediate group and 10 in the third, more challenging group. The first two groups are rather homogeneous and mostly draw from the *De vulgari eloquentia* , while in the third one each work is represented by 2 sentences, and the *Eclogues* are featured only here.

The differences in complexity can be understood in terms of number of nodes, depth, and breadth of the resulting syntactic trees. While a sentence of the first group has a median number of 11 nodes, a median depth of 4 layers and most nodes (not counting the root) tend to be at depth ca. 3, for the second group the same figures are respectively 42, 7 and ca. 4; for the third group they are 46.5, 7.5 and ca. 4.5. The difference is especially marked between the first group and the other two. Besides such quantitative factors, other more qualitative ones, like difficult syntactic structures, contribute to the overall complexity.

As for the second part of our training, which consists of only one phase, textual cohesion substitutes increasing complexity as the main selection criterion: as such, each annotator is assigned, from the work they will respectively take care of, the first 10 sentences which have not been previously annotated. The complexity of the single sentences is thus more variable in this phase, but still well represents the whole corpus. In particular, we use sentences 1-4 and 7-12 from book I of the *De monarchia*, sentences 4-5 and 7-14 from book I of the *De vulgari eloquentia*, sentences 1-10 from the first of Dante's *Eclogues*, and 1-10 from Epistle VII of the *Epistulae*. The *Quaestio de aqua et terra*, of uncertain authorship, is not assigned to any annotator at the moment.

All the selected sentences are priorly syntactically annotated by hand by a UD expert applying language-specific features and subrelations developed for Latin, while lemmas, parts of speech and morphological traits are corrected or enhanced where needed with respect to the CoNLL-U conversion (see Section 2.1). This way, on the one hand a tripartite, scaled gold standard is created for common evaluation, while on the other hand

[10] We add `VerbClass` as a language-specific feature to encode traditional verb conjugations. The value `LatA` indicates the first conjugation, which has thematic vowel 'a'.

[11] From `https://universaldependencies.org/u/pos/ADJ.html`: "*`ADJ` is also used for 'proper adjectives' such as* European *('proper' as in proper nouns, i. e., words that are derived from names but are adjectives rather than nouns).*"

each annotator will be tested on an individual gold standard in the last phase (Section 2.4).

2.3 Tripartite Training Process and Control

The training of the annotators (all with no background in treebanking, but provided with a solid knowledge of Dante's works and academic background in Latin and Italian philology) is split into two main parts: three "training proper" phases (phase 1 to 3), and one further "control" phase (phase 4).

The first part is meant to lay out a common training ground where the annotators can learn the specifics of the UD annotational scheme, and their progress is overseen and periodically reviewed to prompt improvements. In the first phase, the basics of the UD formalism are presented, and the annotators are required to manually annotate a first group of sentences as a way to evaluate their understanding of the UD principles.[12] In the second and third phases, various aspects of the performed annotation get to be discussed and more complex syntactic structures are introduced, each time assigning new, more challenging sentences for an overall evaluation of the annotator's performance and their inter-annotator agreement (see Section 2.2). At every step, the focus is primarily on the syntactic level, since most aspects regarding lemmatisation, parts of speech and morphology are already mostly dealt with during the conversion phase (Section 2.1).

In contrast to the first three phases, the last, control phase is carried out individually for each annotator on separate sets of sentences (see Section 2.2), as a prelude to their actual annotation work.

	Phase 1	Phase 2	Phase 3
EDGES	80%	83%	79%
DEPRELs	84%	92%	91%

Table 1: Inter-annotator agreement.

2.4 Evaluation and Analysis

Table 1 reports the overall inter-annotator agreement (IAA) for each of the first three phases in terms of Fleiss' kappa,[13] with regard to the structure of syntactic trees (EDGES) and the choice of dependency relations (DEPRELs), whereas in Table 2 the correctness of the annotator's analyses are compared for all phases to the gold standard according to label accuracy (LA), unlabelled attachment score (UAS) and labelled attachment score (LAS) (Buchholz and Marsi, 2006). Table 3 presents the macro-average F-measure on the assignment of dependency relations,[14] again for all four phases. Both these scores and the IAA are computed over basic relations only, i. e. disregarding any subrelation (e. g., the dependency relations `obl`, `obl:agent` and `obl:arg` all count as `obl`) so we can focus on the syntactic soundness of the annotations, since more specific subrelations are often related to secondary language-specific, lexical and semantic factors.

For what concerns the IAA, the scores are rather good (always $>75\%$) and, together with the equally positive scores in Table 2, show that the basic principles of UD have been uniformly acquired by all the annotators during the first part of the training, especially the UD scheme of dependency relations. In general, going from phase 1 to phase 3, we notice that all scores are quite stable, and we only observe a slight decrease of the EDGES score in phase 3 which mirrors the noteworthy complexity of the corresponding test sentences (see Section 2.2); the same general decrease shown in Table 2. However, this is more than compensated by generally markedly improved scores for all annotators in phase 4: taking into account the greater variability in sentence complexity, these data show that all annotators have reached a good degree of confidence both with UD's syntactic formalism and with the specific annotation guidelines developed for Latin, which have been constantly updated during this project.

In particular, if we consider only the labelling of single nodes (LA), we register a decided mean improvement in the last phase (89% vs. 79.75%), showing that the annotators have factually improved their assessment of syntactic dependency relations. A similar trend for IAA in the first three phases and quite close scores in Table 3 point to the fact that those cases where annotators disagree are also those for which they have greater uncertainties at the syntactic level; this leads us to con-

[12] `https://universaldependencies.org/guidelines.html`

[13] Our script for calcuating IAA on CoNLL-U files is available at `https://github.com/johnnymoretti/CoNLL-U_Fleiss_Kappa`.

[14] Metrics calculated with MaltEval (Nilsson and Nivre, 2008).

	Phase 1			Phase 2			Phase 3			Phase 4		
	LA	UAS	LAS	LA	UAS	LAS	LA	UAS	LAS	LA	UAS	LAS
Ann1	91%	96%	87%	89%	80%	75%	82%	70%	63%	87%	78%	73%
Ann2	72%	65%	54%	87%	77%	70%	83%	75%	69%	91%	91%	85%
Ann3	78%	83%	72%	86%	84%	78%	83%	76%	70%	92%	87%	82%
Ann4	78%	83%	72%	86%	84%	78%	79%	76%	68%	86%	85%	77%

Table 2: Annotators' performances versus gold standard.

	Phase 1	Phase 2	Phase 3	Phase 4
	F	F	F	F
Ann1	86%	80%	71%	70%
Ann2	54%	69%	73%	86%
Ann3	69%	70%	72%	72%
Ann4	69%	70%	60%	67%

Table 3: Macro-average F-measure on dependency relations with respect to the gold standard.

clude that most errors might stem from the same sources. In particular, while basic core relations, especially for nominals (nsubj, obj), and the choice of the root all score well, we observe most discrepancies, persisting through all phases, with regard to the labelling of clausal dependents, such as advcl and notably clausal complements (ccomp, xcomp). This pairs with minor confusions regarding the labelling and the attachment of connective elements, i. e. both co-ordinating and subordinating conjunctions. These persistent difficulties are reflected by Table 3, which, as a a macro-average that does not take into account the actual frequencies of single dependency relations, has lower scores than LA in Table 2. Considering that the array of syntactic relations in the later phases is much more varied than in the first one, we still observe a quite stable, if not slightly improving, trend.

The decrease of UAS and LAS in the third phase, when compared to the good results of the second phase, has to be expected, as the sentences of phase 3 are chosen to be particularly challenging and in some cases present open problems of syntactic annotation.[15] Despite this, the differences between phase 1 and phase 3 still show a rather stable quality of the annotation from this angle. Then again, the last control phase registers much improved performances also for UAS and LAS, displaying the good level of assurance reached by the annotators at all levels of annotation.

3 Conclusion and Future Work

In this paper, we describe the preliminary steps towards the creation of a UD-compliant treebank of the Latin works by Dante Alighieri. To this end, we create a gold standard and we train and evaluate the work of a team of four annotators by means of a tripartite common set of sentences of increasing complexity annotated by a UD expert, complemented by specific gold standards for each annotator in a final control phase before the actual annotation work takes place.

Besides supporting the objectives of the *Vocabolario Dantesco* project, the development of a treebank based on Dante's Latin works also serves a wider scope, i. e. the inclusion of these latters into the *LiLa Knowledge Base*, which makes distributed linguistic resources for Latin interoperable through the Linked Data paradigm (Passarotti et al., 2020).[16] At the same time, the efforts put into this project will hopefully bring forth some much-needed recommended guidelines for the UD-style annotation of Latin.

The complete annotation of Dante's Latin works will provide the community with a new, manually annotated dataset of higher quality than any automatic system. Table 4 reports LAS scores computed on the sentences of our gold standard and processed with UDPipe using the UD v2.5 models for Latin (Straka and Straková, 2017). The scores clearly show that current models are not good enough to parse the Latin of Dante.

	IT-TB	Perseus	PROIEL
LAS	40.83%	24.93%	29.98%

Table 4: UDPipe scores (based on UD v2.5) for gold standard sentences (all four phases).

[15] See for example the open issue on how to deal with singular subjects and plural copula at https://github.com/UniversalDependencies/docs/issues/714.

[16] https://lila-erc.eu

The addition of Dante's Latin works into the thriving and expanding UD project and the newly acquired possibility to interact with a large number of other Latin texts of different genres and time periods makes us hope for a breakthrough of the world of treebanking into the wider community of the Humanities, which today can benefit from accessing a huge set of connected textual (meta)data like never before.

Acknowledgments

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme via the "LiLa: Linking Latin" project - Grant Agreement No. 769994. The authors also want to thank Prof. Mirko Tavoni (University of Pisa) and the annotators Martina De Laurentis, Federica Favero, Giulia Pedonese and Elena Vagnoni.

References

David Bamman and Gregory Crane. 2011. The ancient Greek and Latin dependency treebanks. In *Language technology for cultural heritage*, pages 79–98. Springer.

Angelo Basile and Federico Sangati. 2016. D (h) ante: A New Set of Tools for XIII Century Italian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2825–2828.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020. A new latin treebank for universal dependencies: Charters between ancient latin and romance languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 933–942.

Hanne Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1):29–65.

Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for Universal dependencies treebank files. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 87–93.

Timo Korkiakangas and Marco Passarotti. 2011. Challenges in Annotating Medieval Latin Charters. *Journal for Language Technology and Computational Linguistics*, 26(2):103–114.

Jürgen Leonhardt. 2009. *Latein. Geschichte einer Weltsprache*. Beck.

Jens Nilsson and Joakim Nivre. 2008. MaltEval: an Evaluation and Visualization Tool for Dependency Parsing. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, LVIII(1):177–212.

Marco Passarotti, 2019. *The Project of the Index Thomisticus Treebank*, pages 299 – 320. De Gruyter Saur, Berlin, Boston.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Mirko Tavoni. 2012. *DanteSearch: il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica*. Università degli Studi di Napoli" L'Orientale", Il Torcoliere-Officine.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: an Overview. In *Treebanks*, pages 5–22. Springer.

Koichi Yasuoka. 2019. Universal Dependencies Treebank of the Four Books in Classical Chinese. In

DADH2019: 10th International Conference of Digital Archives and Digital Humanities, pages 20–28. Digital Archives and Digital Humanities.

Amir Zeldes and Mitchell Abrams. 2018. The Coptic Universal Dependency Treebank. In *Proceedings of the Universal Dependencies Workshop 2018*, pages 192–201, Brussels.